A guide to ancient protein studies

Assessing Ancient Protein Studies: A Guide for Researchers, Reviewers and Editors

Jessica Hendy[1*], Frido Welker[2*], Beatrice Demarchi[3,4], Camilla Speller[4], Christina Warinner[6,7,8], Matthew J. Collins[4,5]

[1] Department of Archaeology, Max-Planck Institute for the Science of Human History, Jena, Germany.

[2] Department of Human Evolution, Max-Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

[3] Department of Life Science and Systems Biology, University of Turin, Italy.

[4] BioArCh, Department of Archaeology, University of York, York, UK.

[5] Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark.

[6] Department of Archaeogenetics, Max-Planck Institute for the Science of Human History, Jena, Germany.

[7] Laboratories of Molecular Anthropology and Microbiome Research, University of Oklahoma, Norman, USA.

[8] Institute for Evolutionary Medicine, University of Zürich, Zürich, Switzerland.

*Equal author contribution. Corresponding authors: hendy@shh.mpg.de and frido_welker@eva.mpg.de

Abstract

Palaeoproteomics is an emerging neologism used to describe the application of mass spectrometry (MS)-based approaches to the study of ancient proteomes. As with palaeogenomics (the study of ancient DNA, aDNA), it intersects evolutionary biology, archaeology and anthropology, with applications ranging from the phylogenetic reconstruction of extinct species to the investigation of past human diets and ancient diseases. However, there is currently no consensus regarding standards for data reporting, data validation measures, or the use of suitable contamination controls in ancient protein studies. Additionally, in contrast to the aDNA

community, no consolidated guidelines have been proposed by which researchers, reviewers and editors can evaluate palaeoproteomics data, in part due to the novelty of the field. Here we present a series of precautions and standards for ancient protein research that can be implemented at each stage of analysis, from sample selection to data interpretation. These guidelines are not intended to impose a narrow or rigid list of authentication criteria, but rather to support good practices in the field and to ensure the generation of robust, reproducible results. As the field grows and methodologies change so too will best practices, and, as with palaeogenomics, it is essential that researchers continue to provide necessary details on how data were generated and authenticated so that the results can be independently and effectively evaluated. We hope that these proposed standards of practice will help to provide a firm foundation for the establishment of palaeoproteomics as a viable and powerful tool for archaeologists, anthropologists, and evolutionary biologists.

Introduction

In 1955, Philip Abelson, nuclear physicist, science writer and then Head of the Geophysical Laboratory at the Carnegie Institution of Washington, published a short paper that laid out what has become, through several cycles of technical advances, the field of palaeoproteomics[1]. In it, he identified the survival of amino acids over geological timescales, the extraordinary stability of proteins, and the role played by hydrolysis in protein survival and diagenesis. In the following decades, new methodologies, such as Edman degradation sequencing and antibody-based protein detection, allowed modern proteins from fresh tissues or cell cultures to be used to study evolutionary questions, providing some of the first insights into the phylogenetic placement of our genus among the great apes[2].

However, with the exception of (chiral) amino acid detection, these technologies were not suitable for the study of ancient proteins. To date, only one ancient protein sequence has been obtained using Edman degradation sequencing technology, a partial osteocalcin sequence recovered from well preserved Moa bone[3]. Antibody-based methods of protein detection have been applied more frequently to ancient samples[4–12], but these methods can only identify epitope

matches and cannot provide further information on amino acid sequences or protein modifications. Moreover, this approach is known to produce false positive reactions due to the generation of epitopes during protein degradation[13], as recently demonstrated in the form of background signal observed following a decade-long ashing experiment at 350°C[14].

Technological advances around the turn of the last millennium changed this perspective with the establishment of new analytical platforms that couple soft-ionization techniques, such as Matrix Assisted Laser Desorption/Ionization (MALDI) and Electrospray Ionization (ESI), with highly sensitive mass spectrometers for the analysis of large biomolecules in solution[15,16]. Currently, the instrument configurations most often used in palaeoproteomics are MALDI Time-of-Flight Mass Spectrometry (MALDI-TOF-MS) and ESI Liquid Chromatography Tandem Mass Spectrometry (ESI-LC-MS/MS), the latter typically linking nano-LC columns to hybrid detectors that combine ion traps/orbitraps, time-of-flights, and/or quadrupoles[17–21].

By detecting peptides and proteins by mass (more specifically, by mass-over-charge, *m/z*), the demands of concentration, purity and reactivity required for Edman degradation sequencing have been overcome, and the challenges of antibody cross-reactivity circumvented. Recently, non-targeted "shotgun sequencing" of ancient peptides using LC-MS/MS has greatly accelerated the field by allowing complex mixtures of thousands of proteins (i.e., proteomes) to be characterized simultaneously from a single sample. Currently, the primary strategy for shotgun analyses is to initially cleave large proteins into shorter peptides, which are then separated by liquid chromatography and analyzed in two subsequent MS events (LC-MS/MS). This approach of identifying proteins from their component parts is termed 'bottom-up' proteomics, in contrast to 'top-down' approaches that analyze native (uncleaved) molecules. In much the same way that ancient (meta)genomics is suited to the relatively short read lengths of most high-throughput DNA sequencing technologies (e.g., Illumina sequencing by synthesis), palaeoproteomics has benefited from MS sequencing technologies that are designed to identify complex protein mixtures from the measurement of many short peptides. This is a critical component of the success of LC-MS/MS methods in palaeoproteomics research; LC-MS/MS does not require intact tertiary, secondary or even primary structure in order to enable protein

identification, and this makes it especially suited to the study of degraded and fragmented proteins, which are known to undergo processes of denaturation and hydrolytic attack through time.

The advent of high-throughput, high-sensitivity mass spectrometry in the past two decades has allowed palaeoproteomics to become increasingly relevant in the fields of archaeology and evolutionary biology. Not only can individual proteins from archaeological and palaeontological contexts be studied, but one can also analyse the complex mixtures of proteins produced by individual organisms (proteomes) or groups of organisms (metaproteomes) found within ancient samples[21–23]. This has facilitated the phylogenetic reconstruction of extant and extinct species[22,24–26], including that of hominins[27], the mechanistic investigation of protein degradation pathways[28], studies of diagenetic and *in vivo* protein post-translational modifications (PTMs)[29–31], the reconstruction of human diet and subsistence patterns[23,32], and the characterization of past human diseases[23,33–36]. The range of tissues and substrates that can be analyzed is similarly broad, including bone, antler, dentine and enamel[21,27,37–39], eggshell[28,40], skin and soft tissues[33,34], dental calculus[41], preserved food remains[42–45], potsherds and ceramic vessels [8,46,47], bindings and glues[47–50], paint binders[51–53], textiles and leather[54,55] parchment[56], mortars[57–59] and soil[60].

Palaeo-Omics in the Hype Cycle

In a recent article, Bösl[61] examined the history of ancient DNA and argued that it follows a 'Hype Cycle' common to many new technologies[62]. Under this model, a 'technological trigger' rapidly generates a '*Peak of Inflated Expectation*' driven by both media and professional interest. However, when the technology fails to match these initial inflated expectations, the discipline falls into a '*Trough of Disillusionment*'. After this retrenchment there is a systematic exploration of the limits of the technology, followed by the implementation of best practices, which allows the discipline or technology to climb up onto the *'Plateau of Productivity*'. Getting to a state of productivity requires the establishment of stringent standards with regard to data acquisition, analysis and reporting. In the 2000s, the field of palaeogenomics accomplished this by

implementing measures to minimize, control and detect exogenous DNA contamination, and by setting a standard of sharing raw DNA sequence data in open repositories[63–66].

Palaeoproteomics, as a relatively young discipline, faces many of the same challenges that the field of ancient DNA did roughly two decades ago. Recent studies of ancient proteins exhibit a wide disparity in data reporting standards, protein authentication measures, and procedures taken to avoid protein contamination (Table 1, Figure 1). As the field expands, it is necessary to develop techniques for distinguishing endogenous and contaminant proteins. This is essential because studies reporting the identification of extraordinary, purportedly ancient proteins without sufficient evidence of authentication have the potential to damage the credibility of this emerging field[67]. Many of the principles put forward in the field of ancient DNA, such as isolation of work areas, the inclusion of negative controls, and the demonstration of appropriate molecular behaviour, provide a useful starting point, but additional measures are necessary. In particular, the conserved nature of proteins compared to DNA renders the authentication of ancient proteins more challenging than that of ancient DNA. For example, within palaeogenomics, the presence of multiple mitochondrial DNA sequences within a single DNA extract can be used to both detect and quantify modern human contamination[68,69]; in contrast, the low amount of intraspecific amino acid sequence variation generally makes it impossible to use protein sequence variation as a criterion by which to detect the presence of multiple contributing individuals of the same species to a single sample. Nevertheless, many concrete steps can be taken in the field, in the laboratory, and during analysis to mitigate the dual challenges posed by contamination and degradation and to improve the identification of endogenous proteins.

Here we present suggested standards and best practices to researchers, reviewers and editors for the generation, analysis, and reporting of ancient protein data in the scientific literature (summarized in Box 1). Our suggestions are intended to integrate with previously established guidelines for modern proteomic studies ([e.g. 70]). We acknowledge that some studies may be unable to adhere to particular procedures due to budgetary, sampling or experimental constraints. However, building on the principles described by Gilbert et al[65], we emphasise that,

at a minimum, researchers must provide details on how data were generated and authenticated, so that others may be able to evaluate ancient protein identifications.

Towards a Standardised Practice of Palaeoproteomics

1. Sample Selection

Choice and critical appraisal of samples is important in ancient protein studies, and several considerations should be kept in mind when designing research. First, some substrates may harbor better potential for preserving endogenous proteins than others. For example, mineralized samples (such as bone, dental calculus and eggshell) may provide a better preservational environment for proteins than other substrates. The presence of a mineral phase can provide protection from degradation driven by external factors, and mineral-organic binding may facilitate the survival of certain peptides by slowing down chemical diagenesis[28]. There may also be differences in protein preservation among different mineralized substrates. For example, peptides may persist longer in eggshell than in bone, in part due to the tighter mineral matrix in eggshell[28,71]. The choice of samples should also be governed by an awareness of the nature and impact of diagenesis. Diagenesis is driven by a complex network of reactions, including chemical degradation (e.g., temperature and age inducing peptide bond hydrolysis and amino acid racemisation) and molecular breakdown driven by environmental factors during burial and storage e.g., microbial decomposition, acid decalcification and water fluctuation[72]. While the interaction of proteins with their substrate may in some cases exceptional preservation (e.g.[28]), these scenarios are generally rare. If a very old sample displays a very well preserved set of peptides, then modern contamination must be considered as a potential factor. For example, the use of consolidants, resins and glues are widespread in museum conservation practice[73]. Such treatments may result in the unintentional introduction of modern proteins, such as animal collagens in glues, plant proteins in natural resins, insect proteins in shellac, and thus researchers should be mindful of the post-excavation history of samples (Figure 1).

For samples with questionable histories or poor preservation, several steps can be taken prior to paleogenomics analysis in order to evaluate protein preservation and to identify potential sources of contamination during burial and storage. One simple approach is to assess the

elemental composition of samples. Amino acid quantification is useful analysis, and concentration and compositional analyses can additionally assess the yield, and in some cases, the character of the preserved proteins[74–76]. In closed systems (e.g. shells) the proportion of free amino acids can reveal the extent of hydrolysis, and this can be complemented by assessment of amino acid racemization, i.e. the increase in levels of D-amino acids[28]. The unusual pattern of D-amino acids from peptidoglycan (found in bacterial cell walls) can also be used to discriminate between chemical and biological decomposition of samples [77]. Peptide mass fingerprinting of bone collagen (i.e., Zooarchaeology by Mass Spectrometry, ZooMS) may also be useful as screening technique to identify the preservation of collagen, as well as the extent of diagenetically induced glutamine deamidation [78,79]. Pyrolysis-GC/MS can be used to detect the presence of amino acids [80], and the absence of amino acids is being used to challenge claims for the detection of proteins in fossil samples [81]. Such assessment and screening should be reported alongside other downstream measures of authentication and interpretation.

2. Laboratory Considerations

Contamination is a central concern in any palaeoproteomics project as it potentially provides false insights into protein composition, phylogeny, and protein modification. Contamination can be introduced at nearly any stage of analysis (Figure 1), but a number of concrete measures can be taken to reduce contamination from modern proteins in the laboratory environment, as well as cross-contamination between ancient samples. Such measures should be described in publications and at a minimum include extraction blanks, the wearing protective clothing including non-latex gloves (latex is a natural product, containing proteins), the use of clean equipment (e.g., washed with bleach solution, alcohol or autoclaved), pure chemical reagents, and no reuse of consumables.

Ancient protein laboratories should make attempts to reduce the presence of proteinaceous material in the background laboratory environment, including keratins from wool, hair, and skin, as well as common protein-based laboratory reagents [82]. Steps to achieve this may include wearing synthetic or cotton clothing (no wool, silk or leather), covering exposed skin on the hands and arms at all times, and using facemasks and hairnets. Additionally, protein-based

laboratory reagents, such as bovine serum albumin (BSA) and chicken lysozyme, should be avoided. If available, the use of a dead air box or positive pressure laminar flow hood is also encouraged in order to provide a sterile or semi-sterile environment where ancient samples can be handled safely.

Cross-contamination from modern proteins can be avoided by completely separating the initial stages of ancient protein research (sampling, extraction, and protein digestion) from other laboratories. The extraction and digestion of ancient proteins should be performed in a location completely separate from experiments working with modern material (e.g., modern food products, cell cultures or tissue studies), and the shared use of reagents and lab consumables between modern and ancient laboratories is strongly discouraged. In the absence of full separation, even if precautions are undertaken to reduce cross-contamination, spurious contamination events can still occur, contributing to doubt when unexpected or extraordinary findings are observed. For example, a recent study reported the identification of two Crimean-Congo hemorrhagic fever virus (CCHFV) peptides in five of six early Iron Age (750-400 BCE) mortuary vessels from Germany [83]. Today, the distribution of this tick-borne virus is limited to the Balkans and parts of Asia and Africa, and little is known about its origins or history, hence making its incidental discovery in Iron Age Germany an extraordinary finding. However, it cannot be overlooked that the research was performed at the University of Texas Medical Branch in Galveston, Texas, a world leader in the study of viral pathogenesis (including CCHFV), nor that the two CCHFV peptides identified are also components of synthetic vectors (reverse genetics vectors pT7-M and pT7-M-ASKA) used to study viral virulence [84]. Hence, to avoid instances of cross contamination, as well as lingering doubts over possible cross-contamination events, we advocate the use of dedicated extraction environments for ancient proteins.

Cross contamination from *ancient* proteins, as opposed to modern, should also be minimized through cleaning of sample processing areas and equipment, by avoiding the reuse of consumables, and by preparing fresh reagents for each batch of sample extractions. Care should also be taken when opening sample tubes to avoid splashing, dripping or aerosol formation, and

samples should not be crowded into tube racks or centrifuges, but rather spaced out with one or more empty wells between samples.

In order to characterize and monitor background laboratory contamination (including the presence of potential contaminants in reagents or consumables), blank extractions should be performed alongside each batch of extractions, and this data should be analyzed, reported and made available in a similar manner to the ancient samples under investigation. This applies to both small-scale experiments on highly valuable samples, as well as to large-scale studies involving hundreds to thousands of samples, such as ZooMS collagen peptide mass fingerprinting of ancient bone fragments [37].

We note that several ancient protein studies report the use of chemical pre-treatments to remove potential surface contamination prior to protein extraction (including ammonium-bicarbonate [85], EDTA [86], or bleaching [71,87,88]. Such steps have proven moderately successful in ancient DNA studies [89–91], but to our knowledge these techniques have not been rigorously tested on ancient protein samples, with the exception of bleaching on carbonate substrates. Research on the effectiveness of protein decontamination techniques on different sample substrates is greatly needed. For example, although surface removal may be effective for some sample types, bone is highly porous and if the sample has been exposed to phases of wetting, or even significant changes of humidity, it is probable that surface contaminants will have migrated below the surface. Additionally, although strong chemical oxidants are potentially useful for removing both surface and subsurface contaminants, they also have the potential to damage surviving endogenous proteins as well, unless the ancient proteins are protected within the intra-crystalline fraction of the mineral matrix [87,88,92].

3. Mass Spectrometry

The current generation of mass spectrometers are powerful, high-throughput and high-performance instruments, and the hardware and operational costs of such systems typically exceed the budget of individual labs. Consequently, most palaeoproteomics research projects utilize mass spectrometers at institutional core facilities, such as those available at many universities, medical schools, and hospitals. These core facilities typically operate at high

volume, running thousands to tens of thousands of samples per year on a single instrument. Because of this, instrument carryover (i.e., the delayed elution of peptides from previous LC runs) is a serious concern, as clients may have little control over how frequently the instrument is cleaned, how often the HPLC columns are changed, or which samples are analyzed before an ancient protein study. As a result, palaeoproteomics projects must build controls into their own research design in order to detect and mitigate potential cross-project and cross-sample carryover events, and instrument parameters, such as the LC column type, MS/MS model, and collision cell type should be described in the manuscript, even when ancient protein extractions are run at an external core facility [93].

Injection blanks or wash buffers should be run before and between each sample during LC-MS/MS analysis in order to clean the column and identify peptide carryover, as peptides persisting in LC columns have the potential to contaminate subsequent samples (Figure 2). The results of these injection blanks (which are distinct from extraction blanks) should be reported in publications, with semi-quantitative analyses of the data [e.g Demarchi et al, Figure 4; 28]. Researchers may need to investigate the extent of carryover in their mass spectrometry set-up before proceeding with sample loading and analysis. In particular, peptides that display strong binding affinities to mineral phases in archaeological/palaeontological material and thus persist through time, may also be those peptides that adhere to LC columns. Therefore, carryover may particularly impact those peptides that we wish to characterise, and thus monitoring the presence of peptides in injection blanks is vital. After flushing the system prior to beginning a palaeoproteomics run, it is recommended to order the experimental samples from most degraded (or oldest or most critical) to least degraded (or more recent or modern controls), and/or in separate batches of MS/MS analyses, in order to further mitigate sample carryover on the instrument. In addition, measuring and standardizing protein concentrations prior to loading LC-MS/MS columns [94] is also advised in order to prevent a highly concentrated sample from contaminating subsequent samples in the loading queue.

Replication is optimal for validating results, in particular for critical samples or for extremely novel results, [e.g. 28]. There are several strategies for validating through replication,

including experimental replication through the complete re-extraction of the same sample in the same laboratory (or, more optimally, in an independent laboratory), or an analytical replication through repeated MS/MS analyses of the same protein extract. We recognize that in cases of small amounts of starting material or very rare or precious specimens, it may not be possible to perform multiple experimental replications. We also realize that replication in independent laboratories might place a significant burden on newly establishing research groups due to the high cost of the analyses and the relatively small number of laboratories currently specializing in ancient protein analysis. Nevertheless, independent replication is a powerful method of validation that should be performed, if at all possible, when reporting novel, extraordinary or unexpected findings. However, it should be noted that in both cases any contamination occurring prior to the introduction of a sample into an ancient protein laboratory will not be identified or resolved by replication (Figure 1).

4. Peptide and Protein Identification

For obtaining peptide identifications from mass spectra, readers may refer to Taylor et al [95]. At a minimum, essential information should be provided on the database used, search tolerances (both MS1 and MS2), fixed and variable protein modifications, false discovery rate (FDR) thresholds, peptide-spectrum matches (PSM) score cut-offs, whether *de novo* and/or error-tolerant matches were allowed, and which algorithm was used to conduct these searches. All peptide identifications, including novel amino acid sequences following Welker et al. [24], should be supported by more than one MS/MS peptide-spectrum match (PSM) Where possible, manual *de novo* verification should be used as a support for novel amino acid sequences.

Additionally, researchers should carefully consider their selection of reference *in silico* databases during data analysis, and should include soil, microbial and/or common contaminant reference databases where appropriate. The failure to select appropriate databases may result in peptide misassignment or even protein misidentification, and taxonomic misassignment is an especially common problem when using small, curated databases. For example, Swiss-Prot, a manually annotated and non-redundant protein sequence database of reviewed protein sequences, contains nearly complete proteomes of model organisms, such as mouse (*Mus musculus*) and

human (*Homo sapiens*), but only partial proteomes of other taxa, such as sheep (*Ovis aries*), goat (*Capra hircus*), cow (*Bos taurus*), and pig (*Sus scrofa*). Eukaryotic peptide searches against Swiss-Prot tend to result in accurate protein assignments but inaccurate taxonomic assignments, whereby conserved peptide sequences are misassigned to model organisms due to an underrepresentation of non-model organisms in the database. For example, in a recent analysis of proteins extracted from a medieval sheep tooth using Swiss-Prot as the search database, it was found that only 20% of the identified eukaryotic proteins were assigned to sheep, while the remaining proteins were misassigned to cattle, human, mouse, pig, and goat [23]. In each case, taxonomic misassignment occurred when the relevant sheep protein was absent from the Swiss-Prot database (Supplementary Table 1). Such database bias is obvious when analyzing archaeological tissues that originate from a single animal, but it poses more serious problems when analyzing metaproteomes, such as those extracted from ceramic residues or dental calculus. Here, multiple species might be expected from a single sample, and database bias must be accounted for in order to avoid the reporting of analytical artifacts and "phantom" taxa.

Because handling of archaeological and palaeontological specimens during excavation and curation provides plenty of opportunities for human protein contamination or cross-contamination from other artefacts (Figure 1), it is recommended to include possible human contaminating proteins in reference *in silico* databases in searches of non-human tissues (for example, animal bones). Ideally this also includes human collagen type I sequences, given this particular protein's resilience to degradation and its presence in the dermis of the skin. Additionally, other skin proteins such as desmoglein-1 (DSG1), dermcidin (DCD) and junctional plakoglobin (JUP), and of course keratins (both from humans and animals) are recurring contaminants. Contaminating keratins may derive from skin and clothing, but also potentially from brushes or other equipment used in sample preparation and conservation. Awareness of the curation history of specific artefacts is therefore essential in the design of search strategies. Future studies focusing on the analysis of mummified skin, ancient furs and textiles will need to address the problem of how to reliably distinguish ancient from modern skin proteins (e.g., through the study of diagenetic protein modifications). Appendix 1 contains a list of commonly encountered contaminants in proteomics laboratories, and consists of the Repository of

Adventitious Proteins cRAP [96]. One should keep in mind that some of the proteins in Appendix 1 may include endogenous proteins depending on the type of sample analyzed (e.g., keratins in furs, egg white proteins in cultural heritage samples, or albumin in bone), and thus care should be taken when interpreting such findings.

Finally, spectral analysis should allow for the types of diagenetic protein modifications typically encountered when dealing with archaeological and palaeontological material, such as glutamine and asparagine deamidation, possibly methionine and tryptophan (di-)oxidation, the formation of pyroglutamic acid, as well as peptide cleavages unrelated to experimentally-derived enzymatic digestion. However the increased dynamic range of instruments mean that low abundance peptides from non-standard tryptic cleavage [97] and variations in both commercial trypsin performance [98] and in-source fragmentation [99] may be mistaken for hydrolysis. When researchers use error-tolerant or *de novo* options to identify protein sequences novel to science, it should be made clear why the results of their particular bioinformatic workflow should be trusted. This can be achieved using statistical parameters and/or actualistic experiments where the correct sequence is known but removed from the searched database [24]. Validation of *de novo* peptide sequences can be achieved by incorporating such modified amino acid sequence into a second round of analysis with a modified sequence database [24,27].

5. Data Interpretation and Authentication

Following data generation, several additional analyses can be performed to further authenticate and affirm the validity of the results. Like DNA, proteins undergo predictable forms of diagenetic alteration over time, so much so that there is an established field of amino acid geochronometry[100], and documentation of diagenetic changes in ancient samples has been suggested as a useful authentication tool. In particular, diagenetically-induced modifications such as glutamine and asparagine deamidation and the presence of non-enzymatic cleavages of individual proteins are expected to occur in ancient samples. The presence of such modifications has been proposed as authenticity markers[23,27,28,32,35,36], but see[101]. Some studies have aimed to contrast such diagenetically-derived protein modifications between different proteins identified in the same sample[27,102], allowing the potential separation of endogenous human proteins from

contaminating human proteins. Knowledge of protein degradation mechanisms and associated modifications is growing especially as a result of the expansion of the field of amino acid racemisation dating (AAR) [for a summary see 92 and references therein], as well as with the growing availability of ancient protein sequence datasets, but understanding these processes in detail will nevertheless require a substantial investment of time and resources in the near future. It should also be noted that some forms of sample preparation can artificially induce these modifications, and thus detailed reporting of experimental methods is essential for correctly interpreting patterns of diagenesis [79].

In addition to documenting and assessing damage patterns, one can further substantiate the taxonomic origin of a protein or peptide by comparing the sequence(s) to a large reference database of genetic or protein sequences. In peptide spectral matching software, such as Mascot (Matrix Science), peptide sequences are inferred by matching produced *m/z* values against a database of *in silico* produced *m/z* values. These *in silico* databases are ideally kept small in order to reduce computational constraints. However, once these sequences have been identified, one can use alignment tools, such as BLAST, to match the identified sequence to a much larger sequence database, containing a larger number of human, animal, and microbial sequences than the first *in silico* database used to first infer the sequence. Therefore, performing this additional step may shed further light on the taxonomic origin of identified peptides. For example, several of the peptides identified in the latest *Brachylophosaurus* protein sequence are identical to their human, mammalian or avian homologues, including those present in ostrich collagen type I [67,103].

Researchers should also be mindful that amino acid modifications can result in modified amino acids having a total mass equalling that of another amino acid. For example, in the case of the whey protein beta-lactoglobulin reported in Warinner et al. [32], it was observed that one of the protein variant sites that distinguishes Bovinae (cattle, yak, and buffalo) from Caprinae (sheep and goats) is an amino acid residue that is aspartic acid in Bovinae, asparagine in sheep, and lysine in goats (Figure 3a). However, the deamidation of asparagine results in its conversion to aspartic acid (Figure 3b) and hence it is not possible to distinguish an unmodified Bovinae residue (D) from a deamidated sheep residue (de. N) at this position (Figure 3c). Only the

identification of an unmodified asparagine (N) or a lysine (K) would therefore allow species discrimination at this site in most situations[27,32]. The presence of diagenetic modifications is particularly challenging for older samples, where deamidation might have converted all surviving endogenous asparagines or glutamines to aspartic acid and glutamic acid respectively, an issue encountered recently for a Middle Pleistocene rhinoceros proteome [102]. Another example of sequence ambiguity is the incomplete fragmentation of a proline-serine peptide bond, which produces a peptide fragment ion isobaric to hydroxyproline-alanine. Cleavage N-terminal to Pro ('the proline effect') is enhanced whilst cleavage C-terminal to proline in MS^2 is depressed[105]. Proline hydroxylation is the most common post-translational modification of collagen, and Ser/Ala is one of the most common substitution pairs; therefore differentiating serine (in effect *hydroxy*alanine) from alanine C-terminal to (hydoxy)proline is especially difficult[30,103].

Even without modification, some amino acids are not distinguishable by mass alone, leading to sequence ambiguity. For example, the amino acids leucine (L) and isoleucine (I) are isobaric and cannot be readily discriminated in typical MS/MS experiments, although it is possible under certain fragmentation conditions (e.g.[104]). When these amino acids are encountered, identification of the residue will be arbitrary and dependent upon the database used. Two approaches that attempt to resolve this issue are currently available[24,26]. If the residues cannot be distinguished, they should ideally be reported as non-discriminated (e.g., I *or* L for all I and L residues) in published sequences.

6. Data Integration and Sharing

Combining proteomic approaches with other biomolecular techniques, where possible, is encouraged, as multiple approaches can be used to supplement or support novel proteomic findings. For example, ancient mtDNA sequences have been used to support palaeoproteomic analyses of hominin taxonomy [27], lipid and proteomic approaches have been used in combination to detect early Bronze Age cereal grains [106 (forthcoming)], and proteomic and isotopic approaches have been used together to identify ancient milk consumption [32].

The sharing of raw and processed mass spectrometric data in public repositories such as the ProteomeXchange [107] is strongly encouraged, and may soon become a requirement in the field. In this era of 'big data' many research communities are mandating the long-term curation of raw datasets in a publicly accessible form, and an updated list of community-recognised repositories is maintained by the journal *Scientific Data* [108]. Accessing and reanalysing raw data is one way that other researchers can test a study's bioinformatic workflow in their own environment. Additionally, archiving allows data to be re-searched in future analyses, and may lead to the identification of additional proteins as reference sequence databases are updated and expanded. This is especially relevant valuable cultural heritage and human/hominin remains, which might not be available for subsequent re-extraction and destructive analysis. Finally, the public sharing of ancient protein data allows such data to be integrated with future biomolecular analysis using different or similar methods, and more generally *"help[s] build rigorous and reliable scientific practices even in the presence of complex experimental challenges*" (Anagnostou et al. [63].

Finally, we call for a critical approach towards the validation of results and data presented in ancient proteins studies. Following Gilbert et al.[65], we suggest that reviewers and editors consider whether the following questions are sufficiently addressed: 1) Are sufficient measures taken to minimize contamination in the laboratory and do data analysis strategies take potential contamination and degradation into consideration?; 2) Is adequate proof of authentic, ancient protein identification presented?; and 3) Is sufficient information presented for independent replication and can the resulting data be examined?

Perspective

Palaeoproteomics holds enormous potential to dramatically expand archaeological, palaeontological and evolutionary research. In light of this promise, we have raised important considerations and have recommended standards for the generation and reporting of ancient protein data. It is our hope that these suggestions will aid non-specialist readers and reviewers of ancient protein publications, as well as assist researchers in palaeoproteomics in improving study

designs. Undoubtedly, with the emergence of new experimental and bioinformatic strategies for characterizing protein degradation and contamination, as well as improved tools for protein validation and authentication, these guidelines will require further refinement and updating in the future. However, it is our hope that the standards of practice presented here will help to provide a firm foundation for the consolidation of palaeoproteomics as a robust tool for evolutionary biology, anthropology and archaeology.

Competing interests

The authors declare no competing financial interests.

Contributions

J.H. and F.W. conceived of and drafted the manuscript. J.H., F.W and C.W. wrote the main text with contributions from the other authors. MJC got all misty-eyed about the early history of ancient proteins.

Acknowledgements

We thank Enrico Cappellini for valuable comments on a previous version of the manuscript. B.D. is grateful to Kirsty Penkman, Jane Thomas-Oates and Julie Wilson for teaching her that being rigorous might make the job slow but the science right, and Roman Fischer for helpful discussion on analytical set-up. This research was supported by the Max Planck Society (J.H., F.W., C.W.) and its Donation Award (to J.H., C.W.), the US National Science Foundation BCS-1516633, BCS-1523264, and BCS-1643318 (to C.W.), the European Research Council under the European Union's Horizon 2020 research and innovation program under grant agreement numbers STG- 678901-FOODTRANSFORMS and STG-677576-HARVEST, the "Rita Levi-Montalcini Young Researchers Programme" (to B.D.), the Wellcome Trust [grant no 108375/Z/15/Z] (to C.S.), a Danish National Research Foundation Niels Bohr Professorship and ERC Investigator Grant 295729-CodeX (M.J.C) and the US National Institutes of Health R01GM089886 (to C.W.).

Appendix 1. FASTA formatted file containing proteins (in)frequently identified as likely contaminants in standard palaeoproteomic research.

Table 1. Reporting of extraction blanks, injection blanks, evidence of protein degradation and MS data reporting in MS/MS-based ancient protein analysis publications. Extraction and injection blanks are marked as present when they are explicitly mentioned in the manuscript; if marked as absent, this does not necessarily suggest that these blanks were in fact not run or analyzed in the experiment, but they are not reported. MALDI-TOF-MS and antibody-based studies are not included. Accession numbers in the final column refer to datasets stored in ProteomeXchange, otherwise the name of other repositories is given; in one case this refers to a university-based ftp page that can be accessed using details provided in the relevant paper. [1]Degradation noted by the presence of smeared gels.

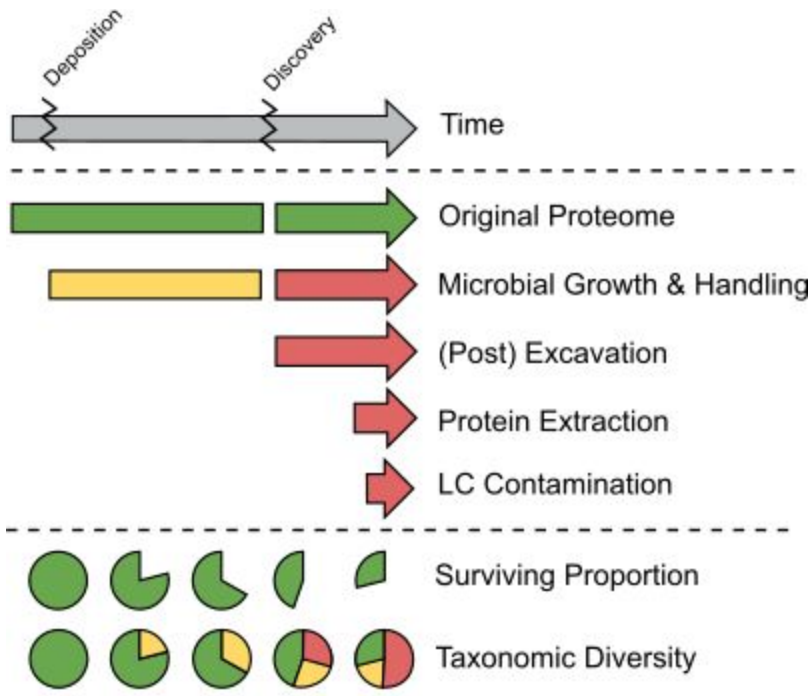


Figure 1. Schematic depiction of ancient proteome compositional changes through time. Initially, the proteome is solely composed of endogenous proteins (green), which may already represent a mixture of taxonomic origins in cases of microbiome samples, food residues, or infected tissues. After deposition, substrates will be rapidly colonized by bacteria and fungi (yellow), some of which might be of interest in future studies. During excavation, curation, and storage, additional contamination can occur, primarily due to human handling and through protein-based consolidants (for example human keratins or animal-based glues; in red). A definitive source of contamination is introduced during sample preparation through the deliberate addition of trypsin, or another protease. Laboratory cross-contamination from both modern and ancient sources can occur during both extraction and LC-MS/MS stages. Throughout the scheme, proteome complexity and protein concentration of the endogenous proteome decrease. Conversely, there is an increase in the proportion of contaminating proteins, both of vertebrate and non-vertebrate origin. Time not to scale. Proportions are used to illustrate general developments and do not necessarily reflect observed frequencies. Modified from Welker [109].

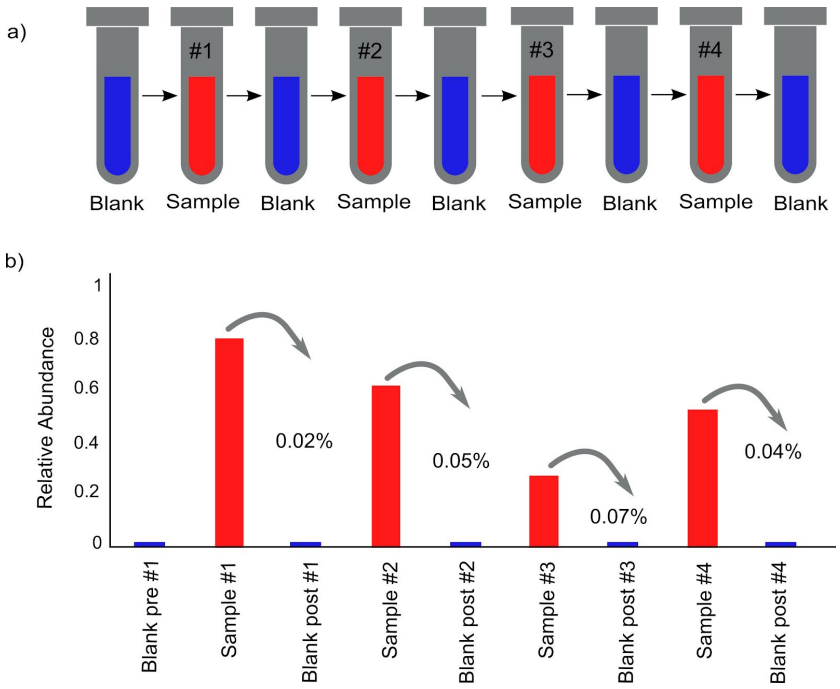


Figure 2. Injection blanks in LC-MS/MS. (**a**) Each sample is preceded and followed by at least one injection blank within the LC column, which (**b**) allows the assessment of peptide carryover between different experiments and samples (following Demarchi et al. [28]). Within this scheme, the extraction blank is analyzed as if representing one of the samples. The blank should show a minimum of a 100-fold reduction in relative abundance of peptides, so that any carry-over for the following samples will be 10,000-fold less.

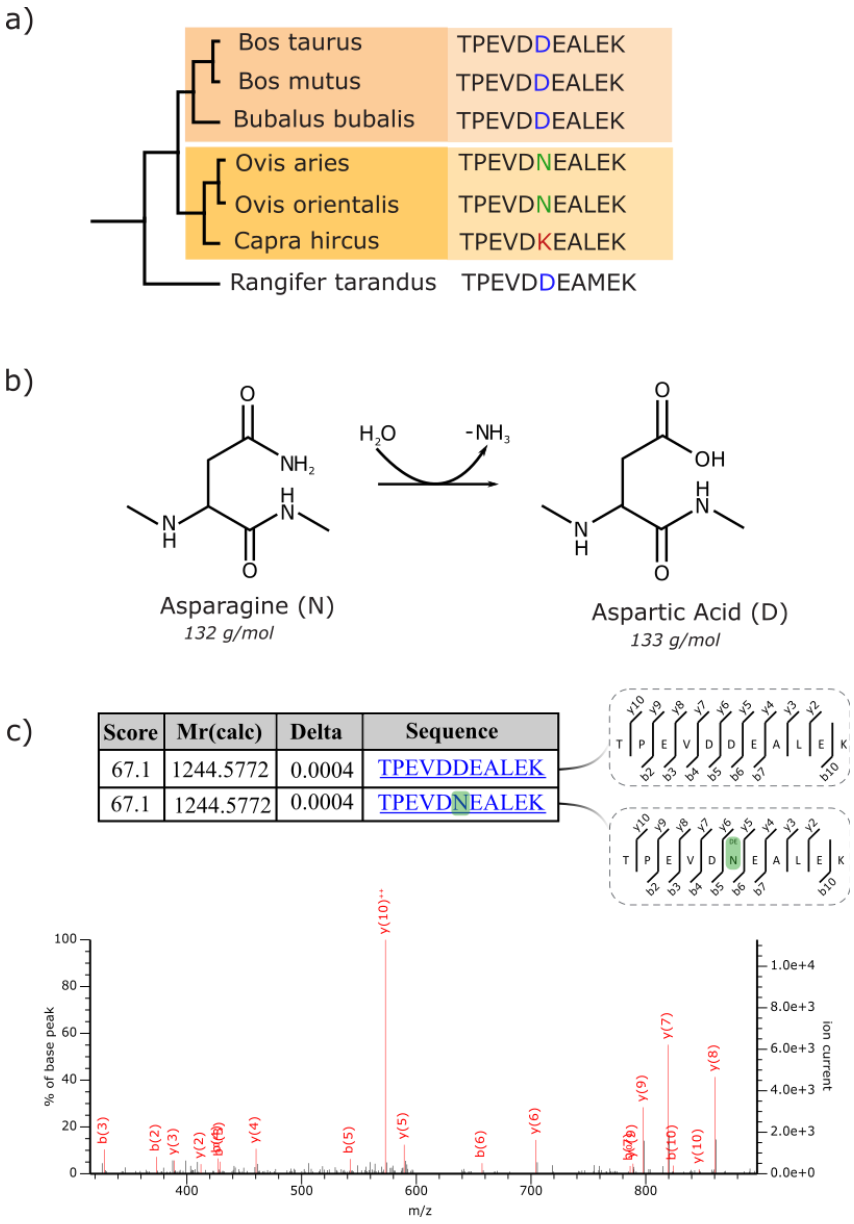
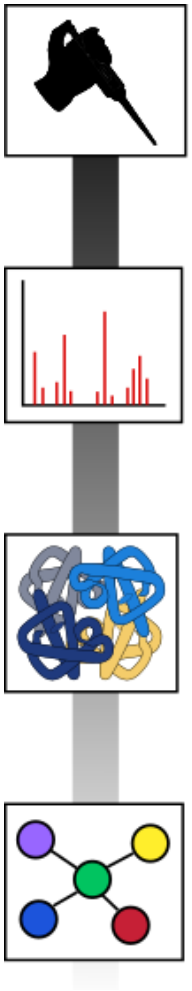


Figure 3. Damage-induced sequence ambiguity affects peptide taxonomic assignment forthe whey protein beta-lactoglobulin. (**a**) An important variant site that distinguishes Bovinae (cattle, yak, and buffalo) from Caprinae (sheep and goats) is an amino acid residue that is aspartic acid (D) in Bovinae, asparagine (N) in sheep, and lysine (K) in goats. However, the deamidation of asparagine results in its conversion to aspartic acid (**b**) and Mascot protein identification software is unable to distinguish an unmodified Bovinae residue (D) from a deamidated sheep residue (de. N) at this position (**c**). Data from [32].



Laboratory Considerations

Use sterilized equipment and pure reagents

Absence of wool, silk, woollen products and latex gloves, and cover exposed skin

Perform blank extractions alongside samples extractions

Separate ancient and modern laboratories and prevent sharing of reagents

Mass Spectrometry

Injection blanks prior to and during LC-MS/MS runs

Perform experimental replications or multiple MS/MS runs for validating results

Peptide and Protein Identification

Include soil, microbial and/or common contaminants in *in-silico* databases

Include diagenetic protein modifications in search strategies

Support novel amino acid sequences with more than one MS/MS PSM

Present statistical parameters or experiments for error-tolerant or de-novo strategies

Data Interpretation and Authentication

Perform BLAST searches of identified peptides in larger reference databases

Perform additional biomolecular techniques to support findings, where applicable

Report the presence of diagenetic PTMs and non-tryptic cleavages

Data Integration and Sharing

Share raw and processed mass spectrometric data on public repositories

Summary Box 1. Some crucial aspects of a paleoproteomics workflow, from extraction to data sharing.

SI Table 1. Demonstration of misleading species assignments in Mascot outputs. Of the top 20 eukaryotic proteins (ranked by score) identified from sheep tooth cementum, only 4 are assigned to sheep. Although the protein identifications themselves are expected for bone/dentine/cementum, misleading species assignments to *Bos taurus, Homo sapiens* and *Mus musculus* are made when the SwissProt database lacks the relevant sheep reference protein.

References

1. Abelson, P. H. Paleobiochemistry: organic constituents of fossils. *Carnegie Institution of Washington, Yearbook, No. 54* (1955).
2. Sarich, V. M. & Wilson, A. C. Immunological time scale for hominid evolution. *Science* **158,** 1200–1203 (1967).
3. Huq, N. L., Tseng, A. & Chapman, G. E. Partial amino acid sequence of osteocalcin from an extinct species of ratite bird. *Biochem. Int.* **21,** 491–496 (1990).
4. de Jong, E. W., Westbroek, P., Westbroek, J. W. & Bruning, J. W. Preservation of antigenic properties of macromolecules over 70 Myr. *Nature* **252,** 63–64 (1974).
5. Westbroek, P. *et al.* Fossil Macromolecules from Cephalopod Shells: Characterization, Immunological Response and Diagenesis. *Paleobiology* **5,** 151–167 (1979).
6. Muyzer, G. *et al.* Preservation of the bone protein osteocalcin in dinosaurs. *Geology* **20,** 871–874 (1992).
7. Schweitzer, M. H., Johnson, C., Zocco, T. G., Horner, J. R. & Starkey, J. R. Preservation of biomolecules in cancellous bone of Tyrannosaurus rex. *J. Vert. Paleontol.* **17,** 349–359 (1997).
8. Craig, O. E. *et al.* Detecting milk proteins in ancient pots. *Nature* **408,** 312 (2000).
9. Collins, M. J. *et al.* Long-term trends in the survival of immunological epitopes entombed in fossil brachiopod skeletons. *Org. Geochem.* **34,** 89–96 (2003).
10. Avci, R. *et al.* Preservation of bone collagen from the late Cretaceous period studied by immunological techniques and atomic force microscopy. *Langmuir* **21,** 3584–3590 (2005).
11. Arslanoglu, J., Schultz, J., Loike, J. & Peterson, K. Immunology and art: using antibody-based techniques to identify proteins and gums in artworks. *J. Biosci.* **35,** 3–10 (2010).
12. Moyer, A. E., Zheng, W. & Schweitzer, M. H. Microscopic and immunohistochemical analyses of the claw of the nesting dinosaur, Citipati osmolskae. *Proc. Biol. Sci.* **283,** (2016).

13. Collins, M. J., Westbroek, P., Muyzer, G. & de Leeuw, J. W. Experimental evidence for condensation reactions between sugars and proteins in carbonate skeletons. *Geochim. Cosmochim. Acta* **56,** 1539–1544 (1992).
14. Moyer, A. E., Zheng, W. & Schweitzer, M. H. Keratin Durability Has Implications for the Fossil Record: Results from a 10 Year Feather Degradation Experiment. *PLoS One* **11,** e0157699 (2016).
15. Twyman, R. *Principles of Proteomics*. (Taylor & Francis, 2004).
16. Lovric, J. *Introducing Proteomics: From Concepts to Sample Separation, Mass Spectrometry and Data Analysis*. (Wiley, 2011).
17. Ostrom, P. H. *et al.* New strategies for characterizing ancient proteins using matrix-assisted laser desorption ionization mass spectrometry. *Geochim. Cosmochim. Acta* **64,** 1043–1050 (2000).
18. Nielsen-Marsh, C. M. *et al.* Sequence preservation of osteocalcin protein and mitochondrial DNA in bison bones older than 55 ka. *Geology* **30,** 1099–1102 (2002).
19. Nielsen-Marsh, C. M. *et al.* Osteocalcin protein sequences of Neanderthals and modern primates. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 4409–4413 (2005).
20. Asara, J. M., Schweitzer, M. H., Freimark, L. M., Phillips, M. & Cantley, L. C. Protein sequences from mastodon and Tyrannosaurus rex revealed by mass spectrometry. *Science* **316,** 280–285 (2007).
21. Cappellini, E. *et al.* Proteomic analysis of a pleistocene mammoth femur reveals more than one hundred ancient bone proteins. *J. Proteome Res.* **11,** 917–926 (2012).
22. Cappellini, E. *et al.* Resolution of the type material of the Asian elephant, Elephas maximus Linnaeus, 1758 (Proboscidea, Elephantidae). *Zool. J. Linn. Soc.* **170,** 222–232 (2014).
23. Warinner, C. *et al.* Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* **46,** 336–344 (2014).
24. Welker, F. *et al.* Ancient proteins resolve the evolutionary history of Darwin/'s South American ungulates. *Nature* **522,** 81–84 (2015).

25. Rybczynski, N. *et al.* Mid-Pliocene warm-period deposits in the High Arctic yield insight into camel evolution. *Nat. Commun.* **4,** 1550 (2013).
26. Cleland, T. P., Schroeter, E. R., Feranec, R. S. & Vashishth, D. Peptide sequences from the first Castoroides ohioensis skull and the utility of old museum collections for palaeoproteomics. *Proc. Biol. Sci.* **283,** (2016).
27. Welker, F. *et al.* Palaeoproteomic evidence identifies archaic hominins associated with the Châtelperronian at the Grotte du Renne. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 11162–11167 (2016).
28. Demarchi, B. *et al.* Protein sequences bound to mineral surfaces persist into deep time. *Elife* **5,** e17092 (2016).
29. Hill, R. C. *et al.* Preserved Proteins from Extinct Bison latifrons Identified by Tandem Mass Spectrometry; Hydroxylysine Glycosides are a Common Feature of Ancient Collagen. *Mol. Cell. Proteomics* **14,** 1946–1958 (2015).
30. Cleland, T. P., Schroeter, E. R. & Schweitzer, M. H. Biologically and diagenetically derived peptide modifications in moa collagens. *Proceedings of the Royal Society B* **282,** 20150015 (2015).
31. Mikšík, I., Sedláková, P., Pataridis, S., Bortolotti, F. & Gottardo, R. Proteins and their modifications in a medieval mummy. *Protein Sci.* (2016). doi:10.1002/pro.3024
32. Warinner, C. *et al.* Direct evidence of milk consumption from ancient human dental calculus. *Sci. Rep.* **4,** 7104 (2014).
33. Corthals, A. *et al.* Detecting the immune system response of a 500 year-old Inca mummy. *PLoS One* **7,** e41244 (2012).
34. Maixner, F. *et al.* Paleoproteomic study of the Iceman's brain tissue. *Cell. Mol. Life Sci.* **70,** 3709–3722 (2013).
35. Kendall, R., Hendy, J., Collins, M. J., Millard, A. R. & Gowland, R. L. Poor preservation of antibodies in archaeological human bone and dentine. *STAR: Science & Technology of*

Archaeological Research **2,** 15–24 (2016).

36. Hendy, J. *et al.* The challenge of identifying tuberculosis proteins in archaeological tissues. *J. Archaeol. Sci.* **66,** 146–153 (2016).
37. Buckley, M., Collins, M. J., Thomas-Oates, J. & Wilson, J. C. Species identification by analysis of bone collagen using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **23,** 3843–3854 (2009).
38. von Holstein, I. C. C. *et al.* Searching for Scandinavians in pre-Viking Scotland: molecular fingerprinting of Early Medieval combs. *J. Archaeol. Sci.* **41,** 1–6 (2014).
39. Stewart, N. A. *et al.* The identification of peptides by nanoLC-MS/MS from human surface tooth enamel following a simple acid etch extraction. *RSC Adv.* **6,** 61673–61679 (2016).
40. Stewart, J. R. M., Allen, R. B., Jones, A. K. G., Penkman, K. E. H. & Collins, M. J. ZooMS: making eggshell visible in the archaeological record. *J. Archaeol. Sci.* **40,** 1797–1804 (2013).
41. Warinner, C., Speller, C. & Collins, M. J. A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370,** 20130376 (2015).
42. Cappellini, E. *et al.* A multidisciplinary study of archaeological grape seeds. *Naturwissenschaften* **97,** 205–217 (2010).
43. Shevchenko, A. *et al.* Proteomics identifies the composition and manufacturing recipe of the 2500-year old sourdough bread from Subeixi cemetery in China. *J. Proteomics* **105,** 363–371 (2014).
44. Yang, Y. *et al.* Proteomics evidence for kefir dairy in Early Bronze Age China. *J. Archaeol. Sci.* **45,** 178–186 (2014).
45. Xie, M. *et al.* Identification of a dairy product in the grass woven basket from Gumugou Cemetery (3800 BP, northwestern China). *Quat. Int.* **426,** 158–165 (2016).
46. Solazzo, C., Fitzhugh, W. W., Rolando, C. & Tokarski, C. Identification of protein remains in

archaeological potsherds by proteomics. *Anal. Chem.* **80,** 4590–4597 (2008).

47. Buckley, M., Melton, N. D. & Montgomery, J. Proteomics analysis of ancient food vessel stitching reveals> 4000-year-old milk protein. *Rapid Commun. Mass Spectrom.* **27,** 531–538 (2013).
48. Dallongeville, S. *et al.* Proteomics applied to the authentication of fish glue: application to a 17 th century artwork sample. *Analyst* **138,** 5357–5364 (2013).
49. Dallongeville, S. *et al.* Identification of animal glue species in artworks using proteomics: application to a 18th century gilt sample. *Anal. Chem.* **83,** 9431–9437 (2011).
50. Solazzo, C. *et al.* Identification of the earliest collagen- and plant-based coatings from Neolithic artefacts (Nahal Hemar cave, Israel). *Sci. Rep.* **6,** 31053 (2016).
51. Hynek, R., Kuckova, S. & Hradilova, J. Matrix‐assisted laser desorption/ionization time‐of‐flight mass spectrometry as a tool for fast identification of protein binders in color layers of paintings. *Rapid Communications in mass spectrometry* (2004). doi:10.1002/rcm.1
52. Tokarski, C., Martin, E., Rolando, C. & Cren-Olivé, C. Identification of proteins in renaissance paintings by proteomics. *Anal. Chem.* **78,** 1494–1502 (2006).
53. Tripković, T. *et al.* Electrospray ionization linear trap quadrupole Orbitrap in analysis of old tempera paintings: application to nineteenth-century Orthodox icons. *Eur. J. Mass Spectrom.* **21,** 679–692 (2015).
54. Brandt, L. Ø. *et al.* Species identification of archaeological skin objects from Danish bogs: comparison between mass spectrometry-based peptide sequencing and microscopy-based methods. *PLoS One* **9,** e106875 (2014).
55. Gong, Y., Li, L., Gong, D., Yin, H. & Zhang, J. Biomolecular Evidence of Silk from 8,500 Years Ago. *PLoS One* **11,** e0168042 (2016).
56. Fiddyment, S. *et al.* Animal origin of 13th-century uterine vellum revealed using noninvasive peptide fingerprinting. *Proceedings of the National Academy of Sciences* **112,** 15066–15071 (2015).

57. Kuckova, S., Hynek, R. & Kodicek, M. Application of peptide mass mapping on proteins in historical mortars. *J. Cult. Herit.* **10,** 244–247 (2009).

58. Krizkova, M. C., Kuckova, S. H., Santrucek, J. & Hynek, R. Peptide mass mapping as an effective tool for historical mortar analysis. *Construction and Building Materials* **50,** 219–225 (2014).

59. Rao, H., Li, B., Yang, Y., Ma, Q. & Wang, C. Proteomic identification of organic additives in the mortars of ancient Chinese wooden buildings. *Anal. Methods* (2014). doi:10.1039/C4AY01766H

60. Oonk, S., Cappellini, E. & Collins, M. J. Soil proteomics: An assessment of its potential for archaeological site interpretation. *Org. Geochem.* **50,** 57–67 (2012).

61. Bösl, E. Zur Wissenschaftsgeschichte der aDNA-Forschung. *NTM* 1–44 (2017).

62. Fenn, J. & Raskino, M. *Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time*. (Harvard Business Press, 2008).

63. Anagnostou, P. *et al.* When data sharing gets close to 100%: what human paleogenetics can teach the open science movement. *PLoS One* **10,** e0121409 (2015).

64. Cooper, A. & Poinar, H. N. Ancient DNA: do it right or not at all. *Science* **289,** 1139–1139 (2000).

65. Gilbert, M. T. P., Bandelt, H.-J., Hofreiter, M. & Barnes, I. Assessing ancient DNA studies. *Trends Ecol. Evol.* **20,** 541–544 (2005).

66. Ginolhac, A., Rasmussen, M., Gilbert, M. T. P., Willerslev, E. & Orlando, L. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* **27,** 2153–2155 (2011).

67. Buckley, M., Warwood, S., van Dongen, B., Kitchener, A. C. & Manning, P. L. A fossil protein chimera; difficulties in discriminating dinosaur peptide sequences from modern cross-contamination. *Proc. Biol. Sci.* **284,** (2017).

68. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23,** 553–559 (2013).

69. Renaud, G., Slon, V., Duggan, A. T. & Kelso, J. Schmutzi: estimation of contamination and

endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16,** 224 (2015).

70. American Society for Biochemistry and Molecular Biology. Instructions to Authors Regarding Required Manuscript Content and Publication Guidelines for Molecular and Cellular Proteomics. (2017). Available at: http://www.mcponline.org/site/misc/ms_guidelines.xhtml. (Accessed: 15th June 2017)
71. Demarchi, B. *et al.* Intra-crystalline protein diagenesis (IcPD) in Patella vulgata. Part I: Isolation and testing of the closed system. *Quat. Geochronol.* **16,** 144–157 (2013).
72. Collins, M. J. *et al.* The survival of organic matter in bone: a review. *Archaeometry* **44,** 383–394 (2002).
73. Schellmann, N. C. Animal glues: a review of their key properties relevant to conservation. *Stud. Conserv.* **52,** 55–66 (2007).
74. Keck, S. & Peters, T. IDENTIFICATION OF PROTEIN-CONTAINING PAINT MEDIA BY QUANTITATIVE AMINO ACID ANALYSIS. *Stud. Conserv.* **14,** 75–82 (1969).
75. Wyckoff, R. W., Wagner, E., Matter, P., 3rd & Doberenz, A. R. COLLAGEN IN FOSSIL BONE. *Proc. Natl. Acad. Sci. U. S. A.* **50,** 215–218 (1963).
76. Brock, F., Geoghegan, V., Thomas, B., Jurkschat, K. & Higham, T. F. G. Analysis of Bone 'Collagen' Extraction Products for Radiocarbon Dating. *Radiocarbon* **55,** 445–463 (2013).
77. Willerslev, E. *et al.* Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science* **317,** 111–114 (2007).
78. van Doorn, N. L., Wilson, J., Hollund, H., Soressi, M. & Collins, M. J. Site-specific deamidation of glutamine: a new marker of bone collagen deterioration. *Rapid Commun. Mass Spectrom.* **26,** 2319–2327 (2012).
79. Simpson, J. P. *et al.* The effects of demineralisation and sampling point variability on the measurement of glutamine deamidation in type I collagen extracted from bone. *J. Archaeol. Sci.* **69,**

29–38 (2016).

80. Stankiewicz, A. B. *et al.* Recognition of Chitin and Proteins in Invertebrate Cuticles Using Analytical Pyrolysis/Gas Chromatography and Pyrolysis/Gas Chromatography/Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **10,** 1747–1757 (1996).
81. Saitta, E. T. *et al.* Low fossilization potential of keratin protein revealed by experimental taphonomy. *Palaeontology* **60,** 547–556 (2017).
82. Hodge, K., Have, S. T., Hutton, L. & Lamond, A. I. Cleaning up the masses: exclusion lists to reduce contamination with HPLC-MS/MS. *J. Proteomics* **88,** 92–103 (2013).
83. Wiktorowicz, C. J., Arnold, B., Wiktorowicz, J. E., Murray, M. L. & Kurosky, A. Hemorrhagic fever virus, human blood, and tissues in Iron Age mortuary vessels. *J. Archaeol. Sci.* **78,** 29–39 (2017).
84. Bergeron, É. *et al.* Recovery of Recombinant Crimean Congo Hemorrhagic Fever Virus Reveals a Function for Non-structural Glycoproteins Cleavage by Furin. *PLoS Pathog.* **11,** e1004879 (2015).
85. Welker, F. *et al.* Variations in glutamine deamidation for a Châtelperronian bone assemblage as measured by peptide mass fingerprinting of collagen. *STAR: Science & Technology of Archaeological Research* **3,** 15–27 (2017).
86. Sawafuji, R. *et al.* Proteomic profiling of archaeological human bone. *Royal Society Open Science* **4,** 161004 (2017).
87. Sykes, G. A., Collins, M. J. & Walton, D. I. The significance of a geochemically isolated intracrystalline organic fraction within biominerals. *Org. Geochem.* **23,** 1059–1065 (1995).
88. Penkman, K. E. H., Kaufman, D. S., Maddy, D. & Collins, M. J. Closed-system behaviour of the intra-crystalline fraction of amino acids in mollusc shells. *Quat. Geochronol.* **3,** 2–25 (2008).
89. Damgaard, P. B. *et al.* Improving access to endogenous DNA in ancient bones and teeth. *Sci. Rep.* **5,** 11184 (2015).
90. Korlević, P. *et al.* Reducing microbial and human contamination in DNA extractions from ancient

bones and teeth. *Biotechniques* **59,** 87–93 (2015).

91. Ginolhac, A. *et al.* Improving the performance of true single molecule sequencing for ancient DNA. *BMC Genomics* **13,** 177 (2012).
92. Demarchi, B. & Collins, M. in *Encyclopedia of Scientific Dating Methods* 1–22 (Springer Netherlands, 2014).
93. Taylor, C. F. *et al.* Guidelines for reporting the use of mass spectrometry in proteomics. *Nat. Biotechnol.* **26,** 860–861 (2008).
94. Noble, J. E. & Bailey, M. J. A. Chapter 8 Quantitation of Protein. *Methods Enzymol.* **463,** 73–95 (2009).
95. Taylor, G. K. & Goodlett, D. R. Rules governing protein identification by mass spectrometry. *Rapid Commun. Mass Spectrom.* **19,** 3420 (2005).
96. The Global Proteome Machine Organization. cRAP protein sequences. (2017). Available at: http://www.thegpm.org/crap/. (Accessed: 1st June 2017)
97. Picotti, P., Aebersold, R. & Domon, B. The implications of proteolytic background for shotgun proteomics. *Mol. Cell. Proteomics* **6,** 1589–1598 (2007).
98. Burkhart, J. M., Schumbrutzki, C., Wortelkamp, S., Sickmann, A. & Zahedi, R. P. Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *J. Proteomics* **75,** 1454–1462 (2012).
99. Kim, J.-S., Monroe, M. E., Camp, D. G., 2nd, Smith, R. D. & Qian, W.-J. In-source fragmentation and the sources of partially tryptic peptides in shotgun proteomics. *J. Proteome Res.* **12,** 910–916 (2013).
100. Penkman, K. & Kaufman, D. Amino acid geochronology: Recent perspectives. *Quat. Geochronol.* 1–2 (2013).
101. Schroeter, E. R. & Cleland, T. P. Glutamine deamidation: an indicator of antiquity, or preservational

quality? *Rapid Commun. Mass Spectrom.* **30,** 251–255 (2016).

102. Welker, F. *et al.* Middle Pleistocene protein sequences from the rhinoceros genus Stephanorhinus and the phylogeny of extant and extinct Middle/Late Pleistocene Rhinocerotidae. *PeerJ* **5,** e3033 (2017).

103. Schroeter, E. R. *et al.* Expansion for the Brachylophosaurus canadensis Collagen I Sequence and Additional Evidence of the Preservation of Cretaceous Protein. *J. Proteome Res.* **16,** 920–932 (2017).

104. Zhokhov, S. S., Kovalyov, S. V., Samgina, T. Y. & Lebedev, A. T. An EThcD-Based Method for Discrimination of Leucine and Isoleucine Residues in Tryptic Peptides. *J. Am. Soc. Mass Spectrom.* (2017). doi:10.1007/s13361-017-1674-3

105. Breci, L. A., Tabb, D. L., Yates, J. R., 3rd & Wysocki, V. H. Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Anal. Chem.* **75,** 1963–1971 (2003).

106. Colonese, A. C. *et al.* New criteria for the molecular identification of cereal grains associated with archaeological artefacts. *Sci. Rep.* (forthcoming).

107. Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32,** 223–226 (2014).

108. Scientific Data. Recommended Data Repositories. (2017). Available at: https://www.nature.com/sdata/policies/repositories. (Accessed: 30th May 2017)

109. Welker, F. The Palaeoproteomic Identification of Pleistocene Hominin Skeletal Remains: Towards a Biological Understanding of the Middle to Upper Palaeolithic Transition. (Max-Planck-Institute for Evolutionary Anthropology, 2017).